Machine Learning 1

Lecture 1.4 - Probability Theory - Bayes Theorem

Erik Bekkers

(Bishop 1.2.0 - 1.2.1)

Slide credits: Rianne van den Berg

Image credit: Kirillm | Getty Images

Probability theory

Probability theory (Bishop)

Provides a consistent framework for the quantification and manipulation of uncertainty.

Uncertainty in pattern recognition

- Noise on measurements.
- Finite size datasets.

Probability theory

Frequentist interpretation

- Probability of event: fraction of times event occurs in experiment

Bayesian approach

- Probability: quantification of plausibility or the strength of the belief of an event.

Random variables

Random variable X

- Stochastic variable sampled from a set of possible outcomes $x \in \mathcal{X}$
- Discrete or continuous
- Probability distribution $\underline{p(X)}$. $\underline{p(x)} \geq 0$, $x \in \mathcal{X}$

Examples of discrete random variables:

- Throwing a dice: $X = \{1, 2, \ldots, 6\}$ $p(x) = \frac{1}{6}$ $\forall x \in \mathcal{X}$
- Flipping a coin: $X = \{heads, tails\}$

$p(X = heads) = \frac{1}{2}$

$p(X = tails) = \frac{1}{2}$

Two discrete random variables (I)

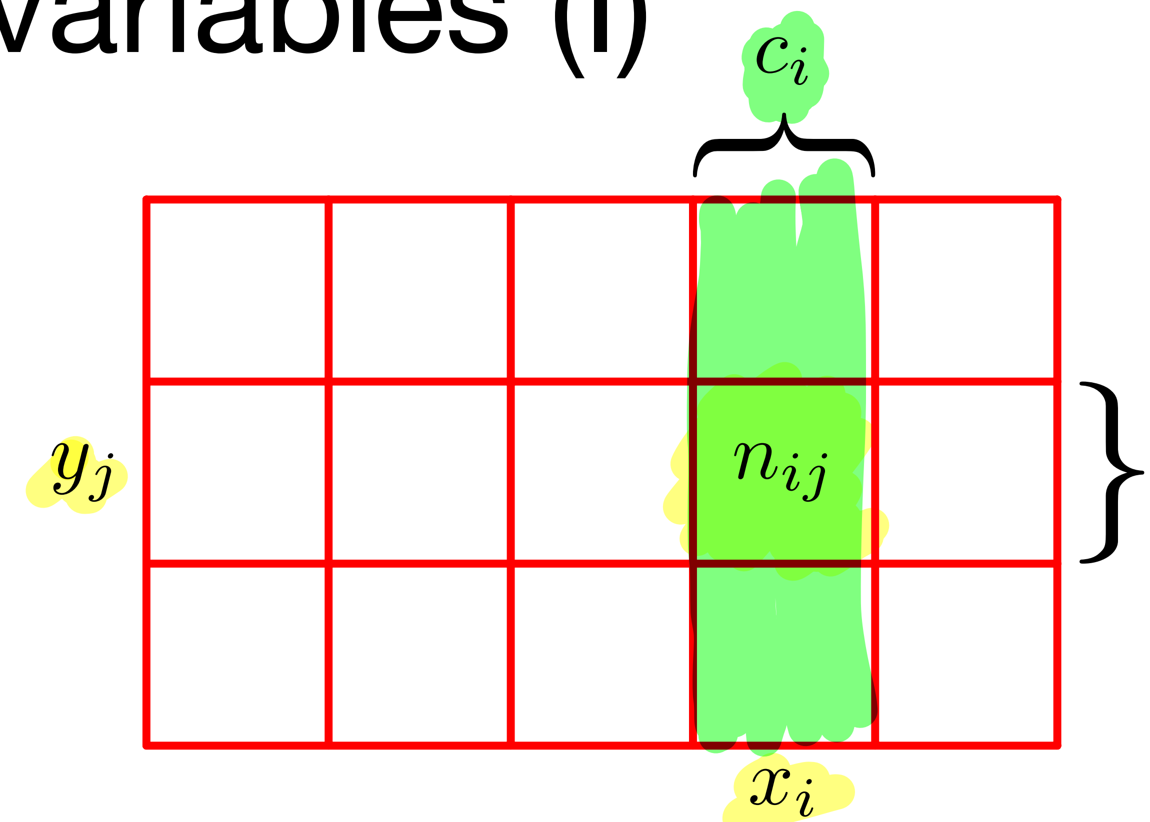


Figure: 2 random variables (Bishop 1.10)

$X = \{x_1, \dots, x_5\}$

2 random variables $Y = \{y_1, y_2, y_3\}$

N trials: sample both *X* and *Y*.

Joint probability

(*) $$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

Marginal probability of X: $p(X = x_i) = c_i / N$ (**)

(**) $$\boxed{c_i = \sum_{j=1}^{3} n_{ij}} \qquad n_{ij} = P(X = x_i, Y = y_j) \cdot N$$

$$P(X = x_i) = \sum_{j=1}^{3} P(X = x_i, Y = y_j)$$

Sum rule of prob

Two discrete random variables (II)

- 2 random variables X, Y

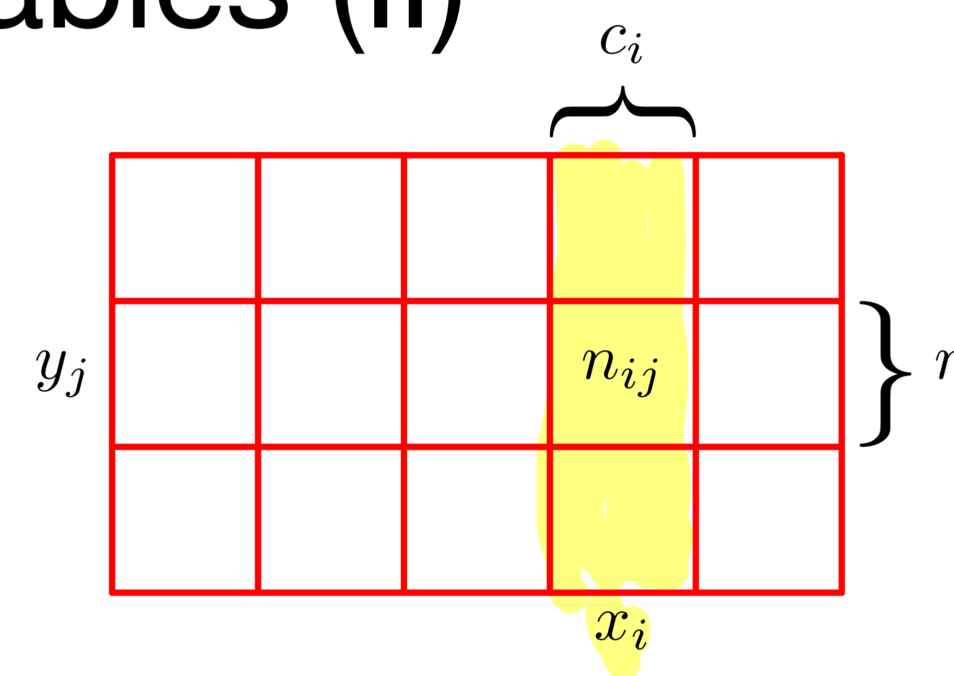


Figure: 2 random variables (Bishop 1.10)

- Conditional probability of Y given X:

$P(Y=y_j \mid X=x_i) = n_{ij}/c_i$

- Remember: $p(X = x_i) = \dfrac{c_i}{N}$

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N} = \frac{P(Y=y_j \mid X=x_i) \cdot c_i}{N}$$

product rule

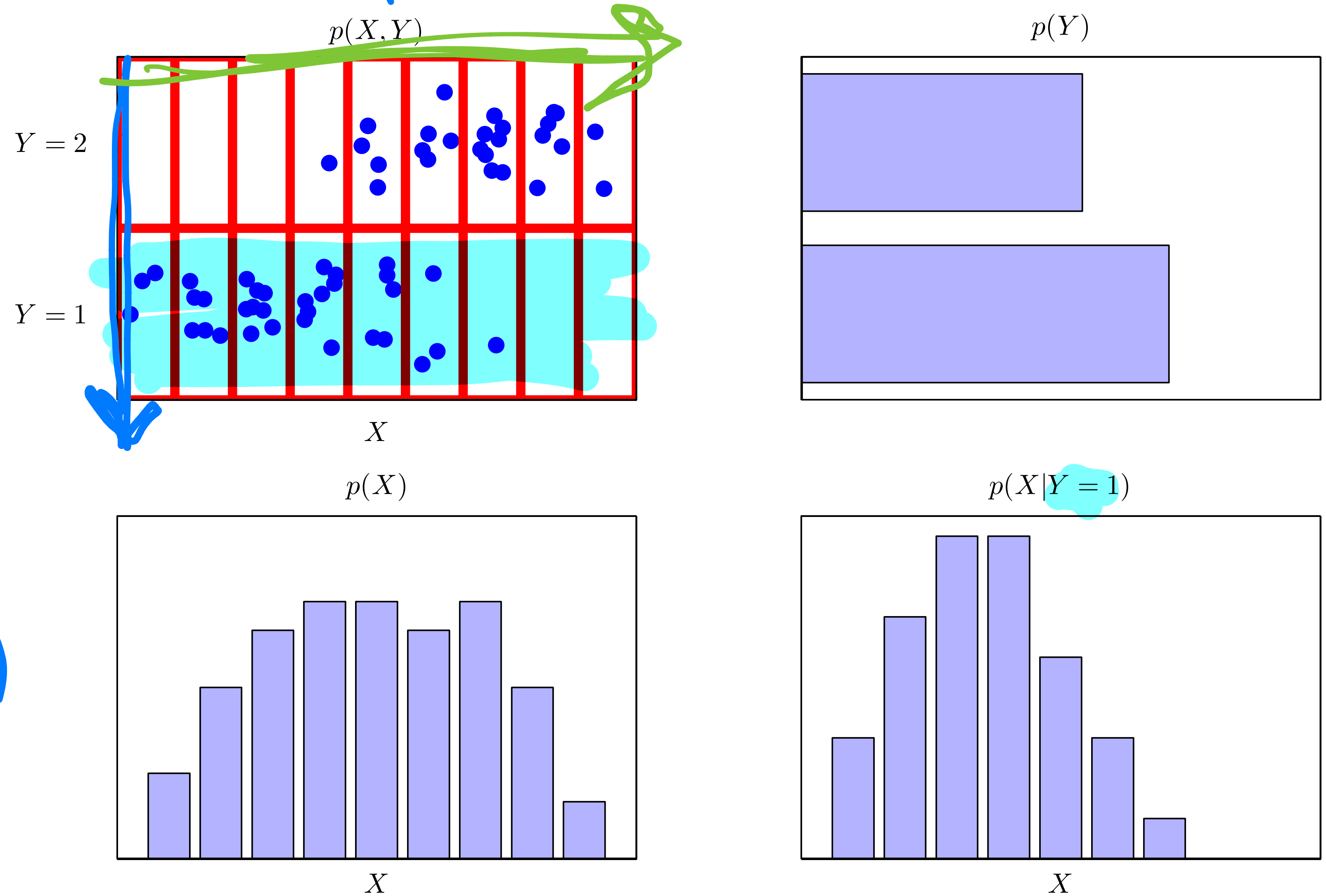
$$P(X=x_i, Y=y_j) = P(Y=y_j \mid X=x_i) \cdot P(X=x_i)$$

Example: Marginal & Conditional distributions

X, Y $\quad D = \{x_i, y_i\}_{i=1}^{60}$

joint distr.

marginal distr $P(X)$



marginal $P(Y) = \sum_{x \in X} p(x, y)$

conditional prob dist $P(y|X)$

Figure: Marginal and conditional distributions (Bishop 1.11)

$\sum_{y \in Y} P(Y = y_i | x_i) = 1$

Continuous Random Variables

- Probability of $x \in \mathbb{R}$ falling in the interval $(x, x + \underline{dx})$ is given by $p(x)\,dx$
- $p(x)$: *probability* *density* over x
- Probability over finite interval $p(\underline{x \in (a, b)}) = \int_a^b p(x)\,dx$
- Positivity: $p(x) \geq 0$
- Normalization: $\int_{-\infty}^{\infty} p(x)\,dx = 1$
- Change of variables *x* = *g(y)*, probabilities in $(x, x + dx)$ must be transformed to $(y, y + dy)$

$$p_x(x)dx = p_y(y)dy \quad \longrightarrow \quad p_y(y) = p_x(x) \left| \frac{dx}{dy} \right|$$

" $|g'(y)|$ "

Continuous Random Variables

Cumulative dist.

$$P(x) = p(X \leq x)$$

$$= \int_{-\infty}^{x} p(\tilde{x})\, d\tilde{x}$$

$$\frac{dP(x)}{dx} = p(x)$$

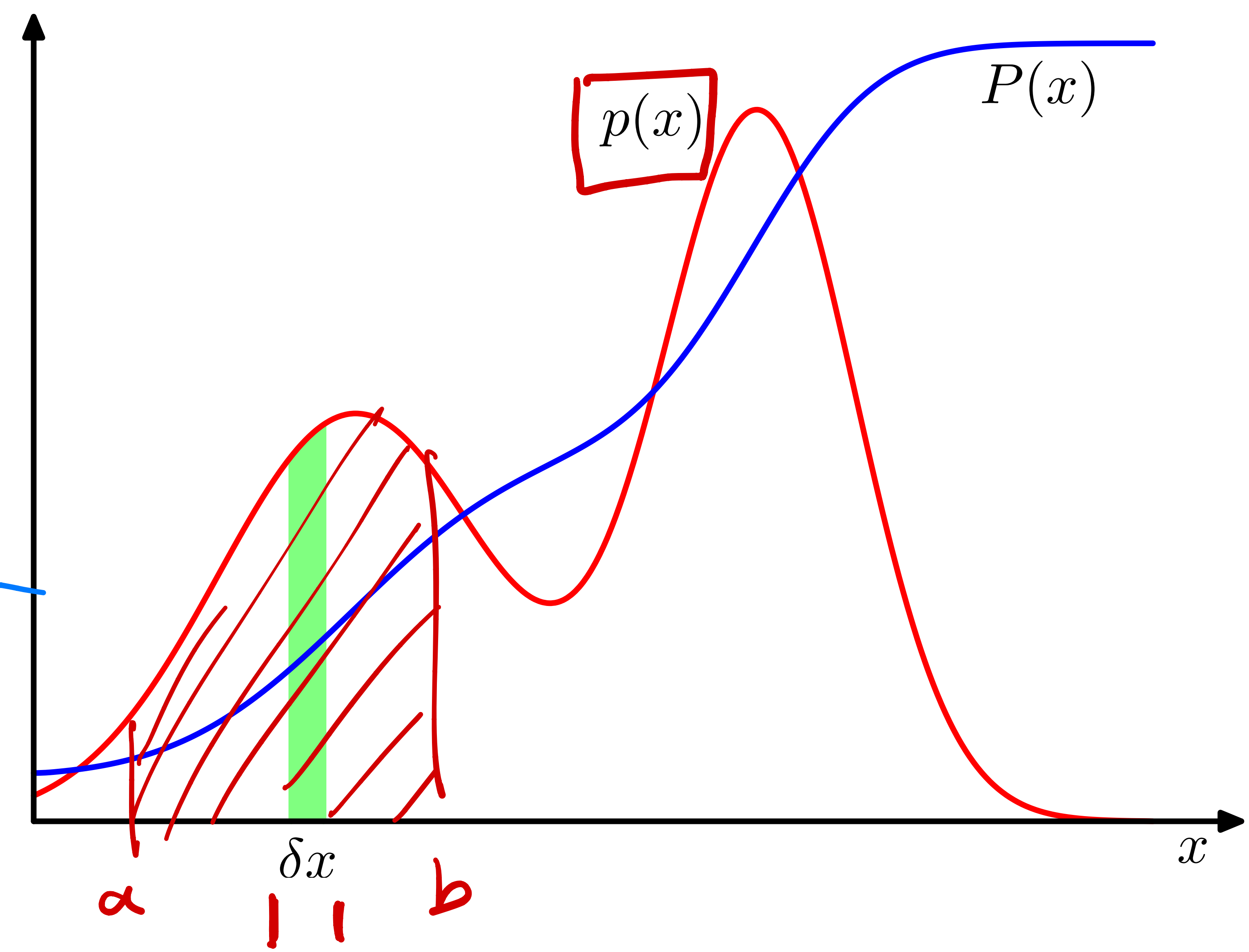


Figure: probability density and cumulative distribution function (Bishop 1.12)

The Rules of Probability Theory

For random variables *X* ∈ $\mathcal{X}$ and *Y* ∈ $\mathcal{Y}$:

	Discrete	Continous
Additivity	$p(X \in A) = \sum_{x \in A} p(x)$	$p(x \in (a,b)) = \int_a^b p(x) dx$
Positivity	$p(x) \geq 0$	$p(x) \geq 0$
Normalization	$\sum_{x \in X} p(x) = 1$	$\int_{\mathcal{X}} p(x) dx = 1$
Sum Rule	$p(x) = \sum_{y \in \mathcal{Y}} p(x, y)$	$p(x) = \int_Y p(x, y) dy$
Product Rule	$p(x, y) = p(x\|y)p(y)$	$p(x, y) = p(x\|y)p(y)$

Bayes Theorem

- Product rule $p(x, y) = p(x|y)p(y)$
- Symmetry property $p(y, x) = p(y|x)\, p(x)$
- Bayes rule ➡ $p(y|x) = \dfrac{p(x|y) \cdot p(y)}{p(x)}$
- Denominator: $\sum_{y \in Y} p(y|x) = 1$

$$\frac{1}{p(x)} \sum_{y \in Y} p(x|y)\, p(y) = 1 \iff p(x) = \sum_{y \in Y} p(x|y)\, p(y)$$

Bayes Theorem

condit. prob. w.r.t. x

function w.r.t. y

Bayes rule

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

- p(y) : the prior probability of Y = y prior / before observing x
- p(y | x) : the posterior probability of Y = y after observing x
- p(x | y) : the likelihood of X = x given Y = y
- p(x) : the evidence for X = x